Sentiment Analysis in Arabic Language

Agenda

- Sentiment analysis
- Important of Sentiment analysis
- General workflow of sentiment analysis
- Challenges in Arabic sentiment analysis
- VADER sentiment analysis

Sentiment analysis

- is also known as opinion mining.
- is a NLP research field that focuses on analyzing people's opinions, sentiments, attitudes, and emotions towards several entities, such as products, services, organizations, issues, events, and topics.
- is a branch of affective computing research that aims to mine opinions from text (but sometimes also images and videos).

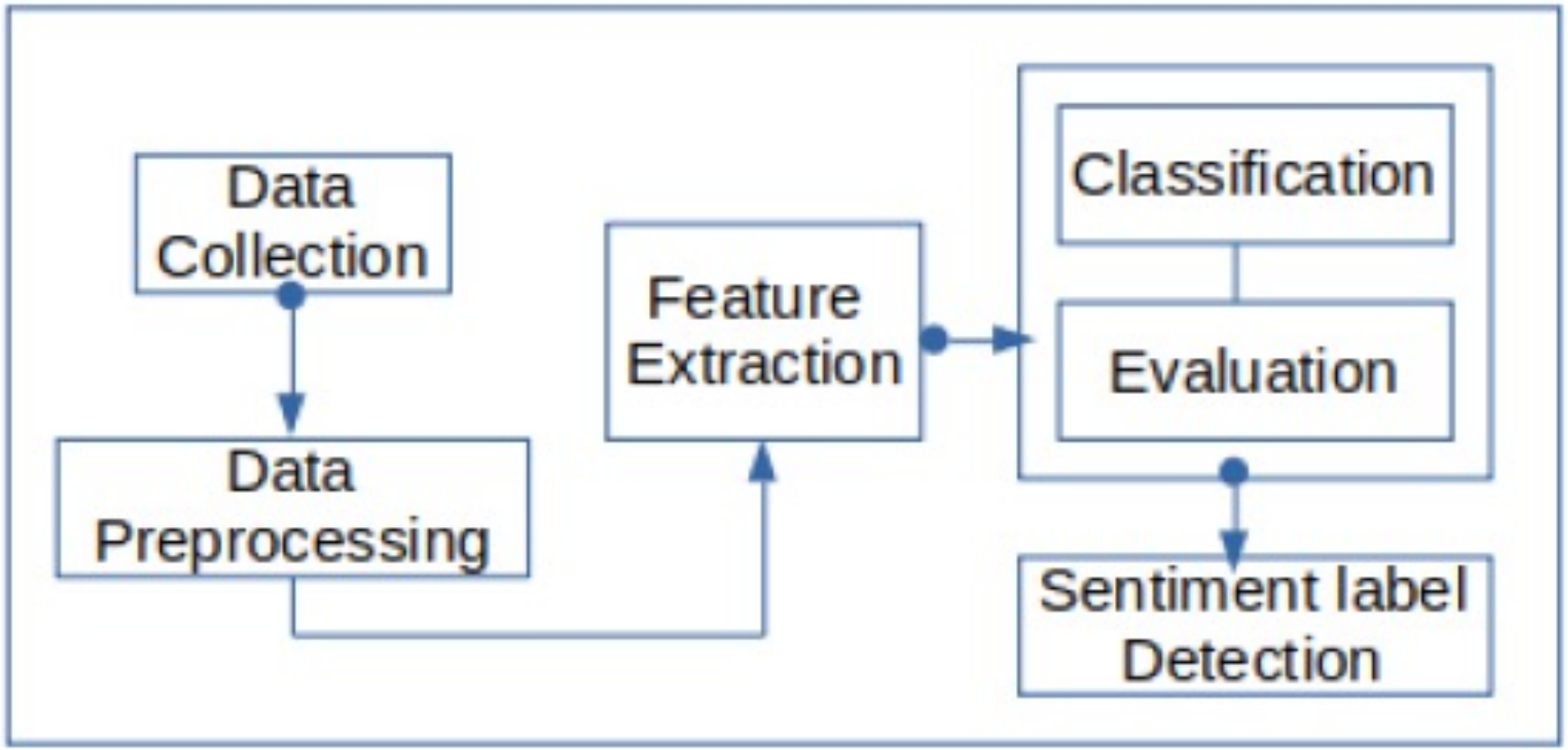
Sentiment analysis

- Sentiment analysis research has mainly been carried out for the English language.
- Although Arabic is ramping up as one of the most used languages on the Internet, only a few studies have focused on Arabic sentiment analysis so far.
- Many studies have tried to use machine translation on English sentiment resources

Important of Sentiment analysis

- It has become a crucial element for decision-makers and business leaders as well as for the public users to understand sentiment and opinions.
- now making important investments in measuring public opinion about their products and services use sentiment analysis tools to extract useful information from unstructured data.

General workflow of sentiment analysis



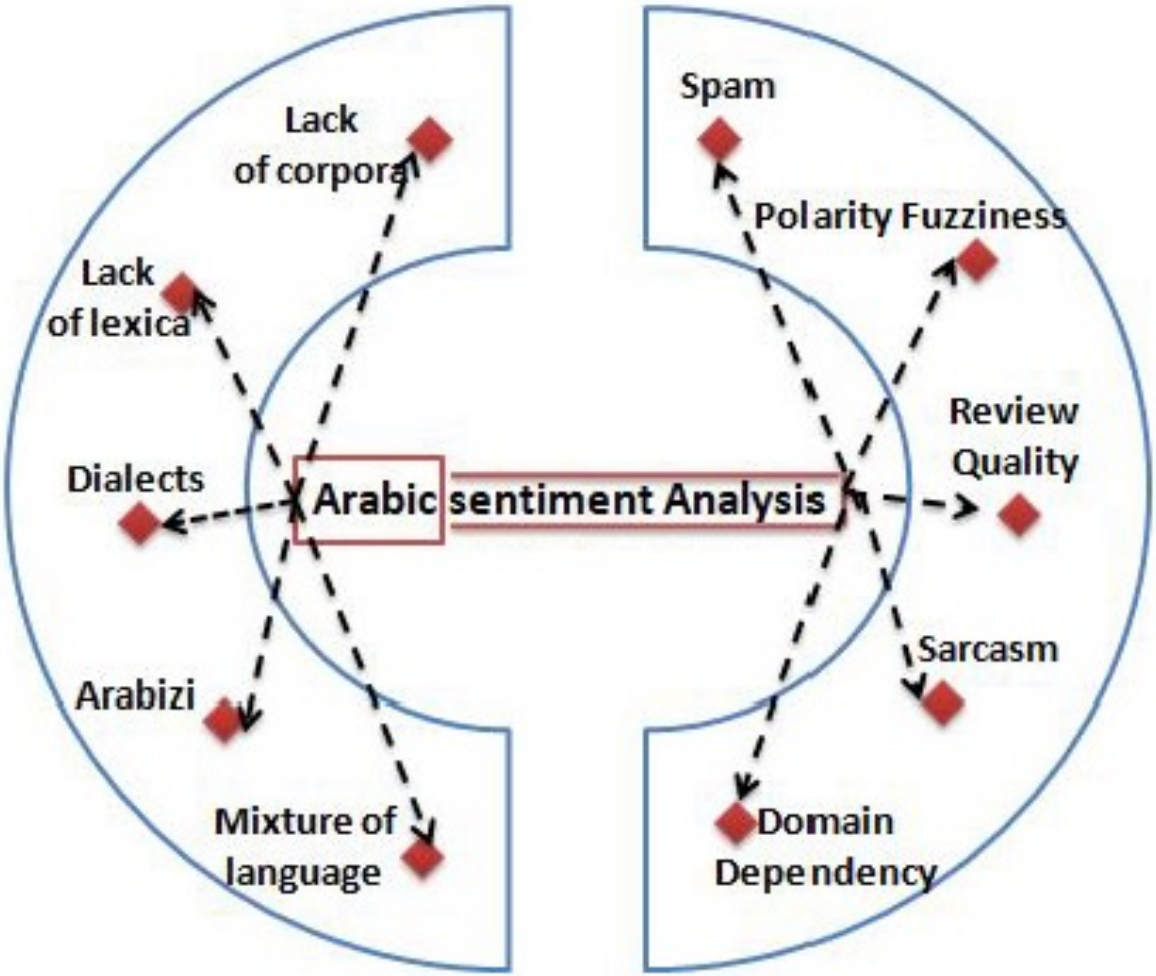
General workflow of sentiment analysis

- Sentiment analysis comprises a multi-step process namely data retrieval, data extraction and selection, data pre-processing, feature extraction, and sentiment classification.
- The **ultimate subtasks of sentiment classification** are
  - polarity classification
  - intensity classification
  - emotion identification

General workflow of sentiment analysis

- **Polarity classification** aims to classify text as positive, negative, or neutral.
- **Intensity classification** seeks to identify the polarity degree (e.g., very positive, positive, fair, negative, very negative).
- **Emotion identification** attempts to identify the specific emotions behind sentiments such as sadness, anger, and fear.

Challenges in Arabic sentiment analysis



Spam

- This activity consists of writing false reviews to promote or to damage the reputation of such service or product.
- It may also be a parasite advertisement which would benefit from the popularity of such webpage or Facebook page, etc.

Spam:


#فرصة_عمل
مطلوب مطلوب مطلوب
مطلوب 50بنت للعمل فورا!!!! بشركه my way

.... و انتى فى بيتك ...مش هيكلفك شئ و لا هيخسرك شئ بالعكس هيكسبك كتير ...
ليكى مشروعك الخاص و دخلك هيوصل ل3000 جنية ....و كمان هدايا و غير كده مكافات
ماليه ... الحقى اشتركى .. هتكونى كسباته معانا اسألى و ان شاء الله هفهمك تشتغلى ازاى و تعملى
ايه ..
لعمل عضوية في شركة ماي واي و التمتع بخصم يصل الي 25% بالاضافة الي هدايا الشركة ..كلمينى ع
رسايل الصفحه ..او سيبي كومنت
Like · Reply · July 22 at 9:31am

Polarity fuzziness:

- Usually, opinion classification systems consider the polarity as positive, negative or neutral.
- It may happen that two humans do not annotate the review in the same way which makes polarity identification difficult.
- For example, while using the emoticon-based approach, a review may contain positive and negative emoticons **at the same time.**
- Also, the review can contain positive (or negative) emoticon and negative (or positive) keyword at the same time.
- Figure 7 illustrates a fuzzy review which contains positive and negative emoticons, but also positive words.

Polarity fuzziness:

Figure 7: A sample of fuzzy review from an Arabic Facebook page

Sarcasm:

- Sarcasm or irony is a form of speech that, in the context of sentiment analysis
- Mostly takes place when the speaker expresses a positive opinion but actually aims to complain about the opinion target .
- Figure 8 illustrates an example of a sarcastic comment posted about an Arabic singer photo.

Sarcasm

Figure 8: A sample of sarcastic review from an Arabic Facebook page

Arabic varieties:

- Arabic is one of the six official languages of the United Nations, and the mother tongue of about 300 million people in 22 different
- Arabic has three main varieties:
  - The **classical Arabic,** also termed Quranic Arabic, is used in religious texts and many old Arabic manuscripts.
  - MSA is the formal language of communication understood by the majority of Arabic speaking people, as it is commonly used in radio, newspapers, and television.
  - Dialectical or colloquial Arabic is used in daily conversation and recently used on both TV and radio.

Arabic orthography:

- MSA texts are written without these marks.
- As a result, a lexical ambiguity problem is created.
- For example, from the same undiacritized word ’قط’, we can derive قِطّ, which means ‘cat’, and ’قَطّ’ which means ’never’.

Lack of corpora:

- A significant factor of an accurate sentiment analysis system is the use of large annotated corpora.
- The accuracy increases proportionally with the quality and the size of the used corpus to train the sentiment classifier.
- **Arabic is poor in terms of corpora** , and the scarcity of Arabic corpora for sentiment analysis is a well-known problem.
- Additionally, the few available corpora are dialectical limited or even free from dialectical content.

Use of dialectal Arabic

- While people in social media express their opinions using their local dialects, the majority of NLP tools are designed to parse MSA.
- Dealing with dialects make the task more complicated because there are no rules, no standard formats either.
- Because of colonization and other historical reasons, many dialectical words are derived from foreign languages such as French and English.
- Some other words are derived from standard Arabic but written differently using Latin letters.
- Table 1 illustrates the gap between MSA and Arabic dialects.

Use of dialectal Arabic

MSA word	Dialectical word	Arabizi	Country	English equivalence	French equivalence
حلو / جميل jameel	حلو	7elew	Lebanon	nice	beau
	حلو	7ilew	Saudi		
	حلو	7low/ hlow	Tunisia		
جدا jiddan	كتير	ktir	Lebanon	Very Wide	trés
	وايد	wayed	Emirate		
	أوي	2awi	Egypt		
	برشا	barcha	Tunisia		
دراجة Darraja	بسكلات	Besklet	Tunisia	Bicycle	Bicyclette
	درا قة	Darraga	Egypt		

Code switching

- Arabic users of social media tend to use Latin characters to represent Arabic words.
- This trend is known as Arabizi.
- The word Arabizi is originated from the portmanteau of Arabic and Englizi. Englizi is how the word English is pronounced in daily Arabic.
- This format is widespread in the Arab world, and most of the new generations use it for code switching, e.g., using Arabic and English or Arabic and French in the same review.

Code switching

- Figure 11 shows a code switching example from Facebook that uses Arabizi and French in the same review.
- The review is “She is beautiful, may God bless her and bless you”.
- The whole review is written using Latin letters.
- However, the first part ‘she is beautiful’ is in French language, and the second part is in Maghrebi dialect.

Figure 11: A sample of code switching (Arabizi and French) from an Arabic Facebook page

Code switching in different alphabets:

- One post can receive multilingual reviews that contain more than a language.
- Figure 12 shows a sample of a mixture of languages (Arabic and French).
- The review is “the best family” the word family is in French, and the word best is in Arabic.
- The review is not only a mixture of two different languages but also of two different alphabets: Arabic and Latin.
- In this case, pre-processing texts by filtering out Latin letters would cause meaning loss and weak classification.

Code switching in different alphabets:

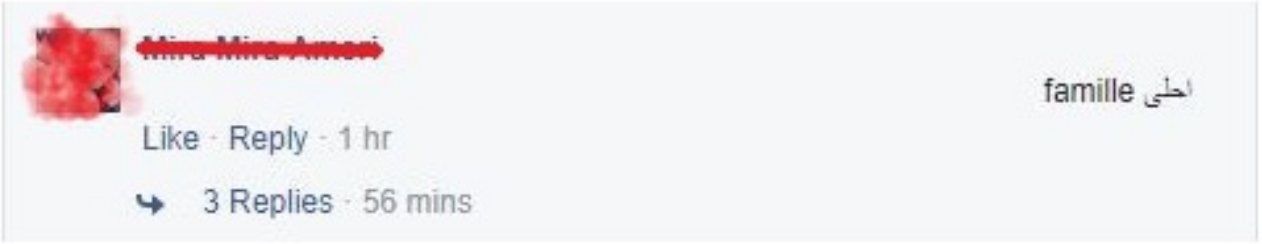


Figure 12: A sample of language mixing from an Arabic Facebook page

VADER sentiment analysis

Python Example

VADER sentiment analysis

- VADER stands for (Valence Aware Dictionary and Sentiment Reasoner).
- Is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media, and works well on texts from other domains.

In [1]:
```
import nltk
#nltk.download('vader_lexicon')
```

In [2]:
```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
SIA = SentimentIntensityAnalyzer()
```

In [3]:
```
a = 'This was a good movie.'
SIA.polarity_scores(a)
```

Out[3]:
```
{'neg': 0.0, 'neu': 0.508, 'pos': 0.492, 'compound': 0.4404}
```

In [4]:
```
a = 'This was the best, most awesome movie EVER MADE!!!'
SIA.polarity_scores(a)
```

Out[4]:
```
{'neg': 0.0, 'neu': 0.425, 'pos': 0.575, 'compound': 0.8877}
```

In [5]:
```
a = 'This was the worst film to ever disgrace the screen.'
SIA.polarity_scores(a)
```

Out[5]:
```
{'neg': 0.477, 'neu': 0.523, 'pos': 0.0, 'compound': -0.8074}
```

References

- Oueslati, O., Cambria, E., HajHmida, M.B. and Ounelli, H., 2020. A review of sentiment analysis research in Arabic language. *Future Generation Computer Systems*, *112*, pp.408-430.